

Incremental Nonlinear System Identification and Adaptive Particle Filtering Using Gaussian Process

Vahid Bastani, *Student, IEEE*, Lucio Marcenaro, *Member, IEEE*, and Carlo S. Regazzoni, *Senior Member, IEEE*

Abstract—An incremental/online state dynamic learning method is proposed for identification of the nonlinear Gaussian state space models. The method embeds the stochastic variational sparse Gaussian process as the probabilistic state dynamic model inside a particle filter framework. Model updating is done at measurement sample rate using stochastic gradient descent based optimisation implemented in the state estimation filtering loop. The performance of the proposed method is compared with state-of-the-art Gaussian process based batch learning methods. Finally, it is shown that the state estimation performance significantly improves due to the online learning of state dynamics.

Index Terms—system identification, incremental learning, online learning, Gaussian process, particle filter, state space model.

I. INTRODUCTION

Bayesian filtering (BF) is the most widespread technique for state estimation in science and engineering. It has been used in many diverse fields including but not limited to signal processing, computer vision, control, robotic and economy. BF requires that the dynamics of the state of the system be known up to some tolerable uncertainty. The fundamental difficulty of BF is to find a correct stochastic process model of the dynamics of the system. Failing to specify a correct and justifiable model will severely impacts the performance of BF and puts it in the risk of undetectable arbitrarily large error.

Linear dynamic model is the commonly used classical model. In this case, Kalman Filter provides efficient and fast solution for BF. However, in the majority of real world applications, the dynamics are nonlinear. Moreover, in the linear models the parameters has to carefully be chosen [1] as well. Particle filtering (PF) is the most flexible form of the BF based on sequential Monte-Carlo that can be applied on nonlinear non-Gaussian dynamic models. Having a correct model in the PF is even more crucial as the PF highly relies on the state dynamic model for sampling process. Filtering under dynamic model uncertainty has been studied in [2], [3] for linear dynamic systems, [4]–[7] for parametric state space models.

In this paper, an incremental/online nonparametric method is proposed for learning nonlinear dynamics in state space model. The Gaussian Process (GP) regression is used here for learning the nonlinear function that models the state dynamic.

V. Bastani, L. Marcenaro and C. S. Regazzoni are with the Department of Electrical, Electronics and Telecommunication Engineering and Naval Architecture (DITEN), University of Genova, 16145, Via All'Opera Pia 11A, Geona, Italy (email: vahid.bastani@ginevra.dibe.unige.it, lucio.marcenaro@unige.it, carlo.regazzoni@unige.it)

Incremental model updating is achieved using the stochastic variational inference of GP. The model updating is integrated inside a PF loop. The proposed method is particularly useful when the measurement data is received in sequence and there is no training data available for learning. Furthermore, when a large number of data is available, it is only practical to process data in sequences or small batches due to the computational resource constraints. One immediate application of learned model is for BF state estimation. This is shown in this paper, where the performance of the PF used in the proposed framework increases gradually since it uses the incrementally learned model for sampling process. However, the learned model can also be used for classification and abnormality detection purposes [8], [9]. Simulating similar data is another application of the learned model which can be used for state prediction as well.

The paper is organized as follows: Section II defines the nonlinear state space model. In Section III the proposed incremental model identification algorithm is presented. In Section IV the performance of the proposed technique is analysed and compared with the state-of-the-art. Finally, Section V concludes the paper.

II. NONLINEAR STATE SPACE MODEL

The state space model (SSM) of a dynamic system is defined using three random processes:

$$\begin{aligned} \mathbf{x}_0 &\sim p_0(\mathbf{x}_0) \\ \mathbf{x}_t|\mathbf{x}_{t-1} &\sim p_f(\mathbf{x}_t|\mathbf{x}_{t-1}) \\ \mathbf{z}_t|\mathbf{x}_t &\sim p_g(\mathbf{z}_t|\mathbf{x}_t), \end{aligned} \quad (1)$$

where \mathbf{x}_t and \mathbf{z}_t are the state and measurement vectors at time t , p_0 is the initial state probability distribution function (PDF), p_f is a conditional probability density function (CPDF) representing the dynamics of the state and p_g is a CPDF representing the measurement process. In a Gaussian nonlinear system the above CPDFs are constructed by:

$$\begin{aligned} \mathbf{x}_t &= f(\mathbf{x}_{t-1}) + \boldsymbol{\omega}_t \\ \mathbf{z}_t &= g(\mathbf{x}_t) + \boldsymbol{\nu}_t, \end{aligned} \quad (2)$$

where f and g are nonlinear functions and $\boldsymbol{\omega}_t$ and $\boldsymbol{\nu}_t$ are zero-mean white Gaussian noises. Conventional state estimation problem considers estimating the posterior of state sequence $\{\mathbf{x}_0, \dots, \mathbf{x}_t\}$ given the measurement sequence $\{\mathbf{z}_1, \dots, \mathbf{z}_t\}$ while all other parameters of the system are known. In BF

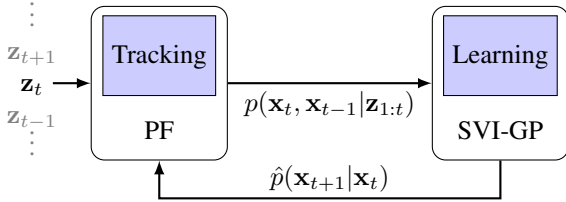


Fig. 1: Simplified diagram of incremental dynamic model identification.

this is achieved by recursively calculating the filtered state posterior:

$$\begin{aligned} p(\mathbf{x}_t | \mathbf{x}_{t-1}, \dots, \mathbf{x}_0, \mathbf{z}_t, \dots, \mathbf{z}_1) &\propto \\ p_f(\mathbf{x}_t | \mathbf{x}_{t-1}) p_g(\mathbf{z}_t | \mathbf{x}_t) p(\mathbf{x}_{t-1} | \mathbf{x}_{t-2}, \dots, \mathbf{x}_0, \mathbf{z}_{t-1}, \dots, \mathbf{z}_1). \end{aligned} \quad (3)$$

This paper deals with the state estimation problem when the dynamic model f is unknown. The goal is to estimate jointly the state sequence and f from measurement sequence. However, the presented technique can be used for estimating g while f is known. Note that when both f and g are unknown the problem is highly ill-posed and can only be attempted with sensible constraints.

III. INCREMENTAL MODEL IDENTIFICATION

Fig. 1 shows a simplified diagram of the proposed incremental identification problem. At the instance t the measurement \mathbf{z}_t is received. The block *Tracker* uses the measurement and the current estimate of the state dynamic model to produce a joint posterior distribution of the current state \mathbf{x}_t and the previous state \mathbf{x}_{t-1} . The posterior is then fed to *Learning* block that uses it for updating the estimate of the state dynamic model.

Due to the nonlinear settings of the problem, the conventional Sequential Importance Resampling (SIR) PF [10] is used here as *Tracker*. In this case then posterior $p(\mathbf{x}_t, \mathbf{x}_{t-1} | \mathbf{z}_{1:t})$ is given as a set of N weighted particles $\{\mathbf{x}_t^{(i)}, \mathbf{x}_{t-1}^{(i)}, \omega^{(i)}\}_{i=1}^N$ with $\omega^{(i)}$ denotes the weight of i^{th} particle. Note that, for the PF algorithm it is only necessary to keep the particle of the current state. However as \mathbf{x}_t and \mathbf{x}_{t-1} are domain and codomain of the function f , it is necessary to jointly estimate both which are then used in *Learning* block. This is achieved by simply keeping particles of previous $t-1$ iteration in the memory. That makes $\mathbf{x}_t^{(i)}$ the filtered state particle and the $\mathbf{x}_{t-1}^{(i)}$ the one-step-lag smoothed state particle.

The *Learning* process incrementally updates the probability model p at each step and provides the updated model to the PF. Stochastic-Variational Sparse Gaussian Process (SVSGP) [11] is used here as *Learning* mechanism. Gaussian Process [12] model is a well established Bayesian nonparametric function regression technique. Its ability for capturing and propagating uncertainties from the training samples to the posterior regression model makes it perfectly fit in the Bayesian framework.

A. Stochastic Variational Gaussian Process

A Gaussian Process (GP) defines a probability distribution over functions $f: \mathcal{X} \rightarrow \mathbb{R}$ such that the marginal distribution of vectorized function values $\Gamma = [f(\mathbf{x}_1), \dots, f(\mathbf{x}_N)]^T$

over any finite subset $\{\mathbf{x}_1, \dots, \mathbf{x}_N\} \subset \mathcal{X}$ be a multivariate Gaussian [12]. A GP, denoted $f(\mathbf{x}) \sim GP(\bar{f}(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$, is characterized by a mean function $\bar{f}(\mathbf{x})$ and a covariance function $k(\mathbf{x}, \mathbf{x}')$ that encodes covariance of two values, $f(\mathbf{x})$ and $f(\mathbf{x}')$.

The GP has widely been applied in Bayesian nonlinear, nonparametric regression problems. Consider training data set $\mathcal{D} = \{X, \mathbf{y}\}$ consists of noisy function values $\mathbf{y} = [y_1, \dots, y_n]^T$ at the set of points $X = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$, where $y_i = f(\mathbf{x}_i) + e$ and e is a white Gaussian noise. The GP regression considers estimating the values of function $\mathbf{f}^* = [f(\mathbf{x}_1^*), \dots, f(\mathbf{x}_M^*)]^T$ at a set of new points $X^* = \{\mathbf{x}_1^*, \dots, \mathbf{x}_M^*\}$ where $f(\mathbf{x})$ has a GP prior. The posterior $p(\mathbf{f}^* | \mathbf{y})$ is a Gaussian with mean vector:

$$\mathbf{K}_{MN}[\mathbf{K}_{NN} + \sigma^2 \mathbf{I}]^{-1} \mathbf{y}. \quad (4)$$

and covariance matrix

$$\mathbf{K}_{MM} - \mathbf{K}_{MN}[\mathbf{K}_{NN} + \sigma^2 \mathbf{I}]^{-1} \mathbf{K}_{NM}. \quad (5)$$

where \mathbf{K}_{MM} , \mathbf{K}_{MN} , \mathbf{K}_{NN} and \mathbf{K}_{NM} are covariance matrices whose elements are $k(\mathbf{x}_i^*, \mathbf{x}_j^*)$, $k(\mathbf{x}_i^*, \mathbf{x}_j)$, $k(\mathbf{x}_i, \mathbf{x}_j)$ and $k(\mathbf{x}_i, \mathbf{x}_j^*)$ respectively. The covariance function and the noise variance control the poster GP. These hyper-parameters are optimized by maximizing the training data marginal log likelihood.

$$\log p(\mathbf{y}) = \log \mathcal{N}(\bar{\mathbf{f}}, \sigma^2 \mathbf{I} + \mathbf{K}_{NN}). \quad (6)$$

The inference in the standard GP has $\mathcal{O}(N^2)$ memory demand and $\mathcal{O}(N^3)$ time complexity. Sparse variational GP [13] reduces complexity by approximating the data set using a variational distribution $q(\mathbf{u}) = \mathcal{N}(\mathbf{m}, \mathbf{S})$ representing the function values over set of inducing points $Z = \{\mathbf{z}_1, \dots, \mathbf{z}_L\}$ that maximize the variational lower-bound of (6):

$$p(\mathbf{y} | Z) > \log \mathcal{N}(\bar{\mathbf{f}}, \sigma^2 \mathbf{I} + \mathbf{K}_{NL} \mathbf{K}_{LL}^{-1} \mathbf{K}_{LN}) \triangleq \mathcal{L} \quad (7)$$

This way the memory and complexity of the inference task will reduce to $\mathcal{O}(NM)$ and $\mathcal{O}(NM^2)$. This can still be prohibitive for *Big Data* problem where N is large. Stochastic Variational Sparse Gaussian Process (SVSGP) [11] proposes another lower bound:

$$\begin{aligned} \mathcal{L} &\geq \mathcal{L}' \triangleq \\ &\sum_{i=1}^N \left\{ \log \mathcal{N}(y_i | \bar{f}_i + \mathbf{k}_i^T \mathbf{K}_{LL}^{-1} \mathbf{m}, \sigma) - \frac{\tilde{k}_{i,i}}{2\sigma} - \frac{\text{tr}(\mathbf{S} \mathbf{\Lambda}_i)}{2} \right\} \\ &- \mathcal{D}_{KL}(q(\mathbf{u}) || p(\mathbf{u})). \end{aligned} \quad (8)$$

where \mathbf{k}_i is the i^{th} column of \mathbf{K}_{LN} , $\mathbf{\Lambda}_i = \sigma^{-1} \mathbf{K}_{LL}^{-1} \mathbf{k}_i \mathbf{k}_i^T \mathbf{K}_{LL}^{-1}$ and $\tilde{k}_{i,i}$ is the i^{th} diagonal of $\mathbf{K}_{NN} - \mathbf{K}_{NL} \mathbf{K}_{LL}^{-1} \mathbf{K}_{LN}$. The difference between \mathcal{L} and \mathcal{L}' is that in the latter the variational distribution parameters are explicit while in the former they are analytically optimized out. However, \mathcal{L}' is written as N terms corresponding to each training data pair. This is the necessary condition for the objective function of stochastic gradient descent (SGD) optimization. The SGD uses approximate gradient from mini-batch in each iteration of gradient descent instead of full gradient calculated on the whole dataset.

The training of SVSGP is done by taking steps in the direction of approximate gradient in each iteration. Since the

approximate gradient is calculated on a subset of training data it is possible to use this in online learning. In online learning the training data is received one by one or in small batches from a supposedly infinite length process.

B. Domain Variable Uncertainty in GP

The standard GP regression assumes training inputs domain are noiseless. This is not the case here as the output of the PF is an estimated joint distribution $\hat{p}(\mathbf{x}_t, \mathbf{x}_{t-1})$ of codomain-domain variables of the GP. domain variable uncertainty in GP has been addressed in [14] for special case of Gaussian i.i.d noise. However, this is not applicable in the problem of this paper as the joint distribution may take any form in the nonlinear dynamics.

A trivial solution is to use particle pairs $\{\mathbf{x}_t^{(i)}, \mathbf{x}_{t-1}^{(i)}\}_{i=1}^N$ as data mini-batches for SVSGP training. However, as the SVGP values all the training data the same and the weights are ignored, this solution is highly inefficient. Alternatively, one may approximate the distribution $\hat{p}(\mathbf{x}_t, \mathbf{x}_{t-1}) = \sum \omega^i \delta(\mathbf{x}_t - \mathbf{x}_t^{(i)}, \mathbf{x}_{t-1} - \mathbf{x}_{t-1}^{(i)})$ with a uniformly weighted particle distribution $\hat{q}(\mathbf{x}_t, \mathbf{x}_{t-1}) = \frac{1}{N} \sum \delta(\mathbf{x}_t - \tilde{\mathbf{x}}_t^{(i)}, \mathbf{x}_{t-1} - \tilde{\mathbf{x}}_{t-1}^{(i)})$ and use the equally weighted particles set $\{\tilde{\mathbf{x}}_t^{(i)}, \tilde{\mathbf{x}}_{t-1}^{(i)}\}$ as mini-batches for GP training. \hat{q} can be optimized by minimizing the KL divergence:

$$KL(\hat{p}||\hat{q}) = \sum \omega^{(i)} \log \frac{N\omega^{(i)}}{\eta_i} \quad (9)$$

subject to $\sum \eta_i = N$ and $\eta_i \in \mathbb{N}$, where η_i is the number of elements in $\{\tilde{\mathbf{x}}_t^{(j)}, \tilde{\mathbf{x}}_{t-1}^{(j)}\}_{j=1}^N$ that are equal to $(\mathbf{x}_t^{(i)}, \mathbf{x}_{t-1}^{(i)})$. It is easy to verify that the η_i that solves (9) have to be approximately proportional to $\omega^{(i)}$. In fact, solving for \hat{q} is exactly equivalent to resampling process in the PF for particle degeneracy mitigation [10].

Resampling replicates particles with larger weights and removes low weight particles. Using resampled particles for GP training artificially incorporates their weights since the contribution of each particle get multiplied proportional to its weights due to the summation in GP objective function (8).

C. The Algorithm

Algorithm 1 shows one iteration of the proposed method. $\hat{\sigma}_t$, $\hat{\theta}_t$, $\hat{\mathbf{m}}_t$ and $\hat{\mathbf{S}}_t$ denote estimated dynamic noise variance, parameter of GP kernel, mean of q and covariance of q respectively after t^{th} measurement. The gradient descend step $GD(\dots)$ is done by in the standard way.

Algorithm 1 An iteration of incremental model identification

Input: $\mathbf{z}_t, \{\mathbf{x}_{t-1}^{(i)}, \mathbf{x}_{t-2}^{(i)}, \omega^{(i)}\}_{i=1}^N, \hat{\sigma}_{t-1}, \hat{\theta}_{t-1}, \hat{\mathbf{m}}_{t-1}, \hat{\mathbf{S}}_{t-1}$

Output: $\{\mathbf{x}_t^{(i)}, \mathbf{x}_{t-1}^{(i)}, \omega^{(i)}\}_{i=1}^N, \hat{\sigma}_t, \hat{\theta}_t, \hat{\mathbf{m}}_t, \hat{\mathbf{S}}_t$

- 1) Optionally resample $\{\mathbf{x}_{t-1}^{(i)}, \mathbf{x}_{t-2}^{(i)}, \omega^{(i)}\}_{i=1}^N$ to avoid degeneracy.
- 2) Sample $\mathbf{x}_t^{(i)} \sim \hat{p}_f(\mathbf{x}_t | \mathbf{x}_{t-1}^{(i)})$ for $i = 1, \dots, N$.
- 3) Let $\omega^{(i)} = \omega^{(i)} p_g(\mathbf{z}_t | \mathbf{x}_t^{(i)})$ for $i = 1, \dots, N$.
- 4) Resample $\{\mathbf{x}_t^{(i)}, \mathbf{x}_{t-1}^{(i)}, \omega^{(i)}\}_{i=1}^N$ to $\{\tilde{\mathbf{x}}_t^{(j)}, \tilde{\mathbf{x}}_{t-1}^{(j)}\}_{j=1}^N$ to minimize (9).
- 5) Calculate \mathcal{L}' for $\{\tilde{\mathbf{x}}_t^{(j)}, \tilde{\mathbf{x}}_{t-1}^{(j)}\}_{j=1}^N$ from (8).
- 6) Calculate gradient $\nabla \mathcal{L}' = [\frac{\partial \mathcal{L}'}{\partial \hat{\sigma}_{t-1}}, \frac{\partial \mathcal{L}'}{\partial \hat{\theta}_{t-1}}, \frac{\partial \mathcal{L}'}{\partial \hat{\mathbf{m}}_{t-1}}, \frac{\partial \mathcal{L}'}{\partial \hat{\mathbf{S}}_{t-1}}]$ for $\{\tilde{\mathbf{x}}_t^{(j)}, \tilde{\mathbf{x}}_{t-1}^{(j)}\}_{j=1}^N$ [11].
- 7) Calculate new parameters using gradient descend: $\hat{\sigma}_t, \hat{\theta}_t, \hat{\mathbf{m}}_t, \hat{\mathbf{S}}_t \leftarrow GD(\mathcal{L}', \nabla \mathcal{L}', \hat{\sigma}_{t-1}, \hat{\theta}_{t-1}, \hat{\mathbf{m}}_{t-1}, \hat{\mathbf{S}}_{t-1})$

IV. EVALUATION

A. Comparison

The performance of the proposed method is compared with GP-SSM [15] and GP-NARX [16] which are both GP-based. Unlike proposed method, these two methods are batch based that is working on full training data. It should be noted that [15] also proposes a stochastic variational inference and discusses possible online application, but it is left without elaboration. The same evaluation setup in [15] is used here for comparison. The algorithms applied on the samples of a nonlinear dynamic model defined by $p(x_t | x_{t-1}) = \mathcal{N}(f(x_{t-1}), 1)$ and $p(z_t | x_t) = \mathcal{N}(x_t, 1)$ where

$$f(x) = \begin{cases} x + 1 & x < 4, \\ -4x + 21 & x \geq 4. \end{cases} \quad (10)$$

Table I compares the performances of the proposed method with the state-of-the-art. The methods are trained with a sequence of 500 samples then they are tested with another sequence of 10^4 samples. The Matérn kernel is used for all GP based algorithms. Fig. 2 shows the test function and the function learned by the proposed method. The performance metrics are the Mean Squared Error (MSE) between the test samples and the predictions and the Mean Log Likelihood (MLL) of the test samples given the trained model $p(x_t^{\text{test}} | x_{t-1}^{\text{test}})$. As the Table I shows, despite the proposed method is incremental/online, its performance is comparable to the state-of-the-art. GP-SSM. The MSE is slightly higher than the GP-SSM while the MLL is improved a little.

TABLE I: Learning performance comparison

Method	Test MSE	Test MLL
Proposed (incremental)	1.17	-1.56
SSM-GP (batch)	1.15	-1.61
GP-NARX (batch)	1.46	-1.90

B. Performance

The incremental learning performance of proposed method is evaluated using simulated nonlinear dynamic models given as

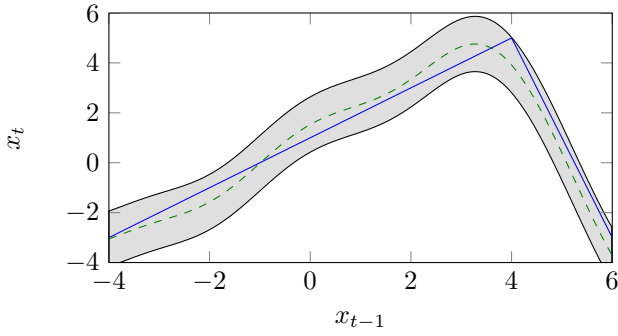


Fig. 2: The test function and the output of the GP trained with proposed algorithm.

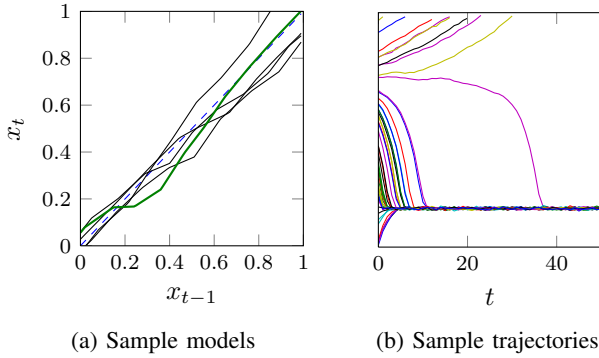


Fig. 3: Some samples of simulated dynamic models (a) and trajectories (b) used for evaluation

$p(x_t|x_{t-1}) = \mathcal{N}(f(x_{t-1}), 10^{-2})$ and $p(z_t|x_t) = \mathcal{N}(x_t, 10^{-3})$ where

$$f(x) = x + \begin{cases} \frac{b_1-b_0}{a_1-a_0}(x-a_0) & a_0 \leq x < a_1 \\ \vdots & \vdots \\ \frac{b_n-b_{n-1}}{a_n-a_{n-1}}(x-a_{n-1}) & a_{n-1} \leq x < a_n \end{cases} \quad (11)$$

with $(a_i - a_{i-1}) \sim \mathcal{U}(0.08, 0.15)$, $(b_i - b_{i-1}) \sim \mathcal{N}(0, 10^{-3})$ and $n = 20$. Unlike (10), (11) generates smooth trajectories which are more realistic as systems are usually constrained by energy. 50 random functions are generated from (11) by sampling a_i and b_i . Five samples of such function are shown in Fig. 3a. Using each random function 50 trajectories are simulated with $p(x_0) = \mathcal{U}(0, 1)$. The models are producing diverse trajectory shapes. Fig. 3b shows sample trajectories generated by the highlighted function in Fig. 3a.

The proposed method is applied on each of the 50 models separately. The trajectories of the model are sequentially fed into the algorithm. The range of measurement is assumed to be $[0, 1]$. If the trajectory goes beyond the scope, it is truncated and no further processing is applied on that. The tracking performance of the PF is recorded for every trajectory in terms of the MSE between the ground truth trajectory and the estimation by PF, i.e. $MSE_i = 10 \log \sum (\hat{x}_t^i - x_t^i)^2 / T$ for i^{th} trajectory. It is expected that over the time the tracker performance improves as the algorithm updates the learned dynamic model with each measurement. Fig 4 shows the scatter plot and the KNN average (red line) of MSE_i versus the

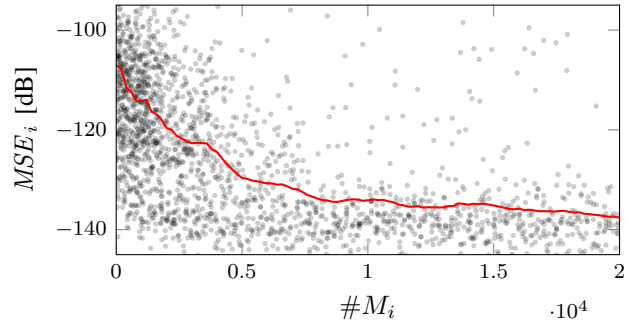


Fig. 4: Tracker MSE with respect to number of received measurements.

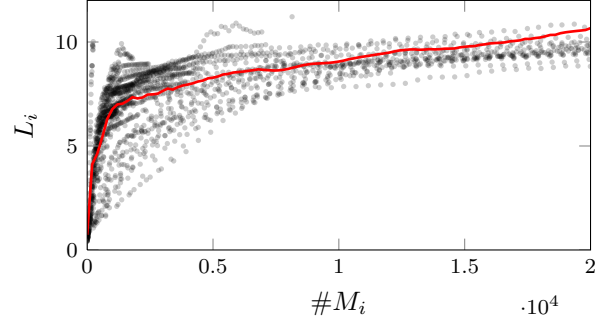


Fig. 5: Ground-truth function likelihood with respect to number of received measurements.

total number of measurements in all the trajectories received before i , i.e. $\#M_i = \sum_{j=1}^{i-1} |\{z_0^j, \dots\}|$. It is clear from Fig. 4 that by incrementally learning the true dynamic model the performance of PF significantly improved over 25dB.

Let $L_i = p([f(x_1^*), \dots, f(x_N^*)]|\theta_i)$ be the likelihood of the ground truth function evaluated on sample point x_1^*, \dots, x_N^* given the learned GP model θ_i up to processing of i^{th} trajectory. The L_i is a relative indication of the closeness of the learned function to the ground truth function. It is used for evaluating the quality of the incremental learning algorithm with $N = 10^4$ and x_1^*, \dots, x_N^* uniformly distributed over $[0, 1]$. Fig 5 shows the scatter plot of L_i versus $\#M_i$ as well as the KNN average of the values. The empirical convergence of the proposed method is relatively fast. It averagely converges with less than 2000 measurement as shown by Fig. 5.

V. CONCLUSION

A sparse Gaussian process based incremental nonparametric system identification method for nonlinear state space models is proposed in this paper. The method is able to update an estimate of the with every measurements from the system. The grid inducing point positioning of the proposed method is particularly limits its usage in high dimensions since lots of the inducing points will placed in the regions the may not visited by any data. Another limitation of the proposed method is that due to the underlying assumption that the dynamics can be model by function. This will fail when the dynamics is multi modal, i.e. depending on some latent effects the dynamic model changes. In future these limitations have to be addressed.

REFERENCES

- [1] H. Heffes, "The effect of erroneous models on the kalman filter response," *IEEE Transactions on Automatic Control*, vol. 11, no. 3, pp. 541–543, Jul 1966.
- [2] T. Ardeshiri, E. zkan, U. Orguner, and F. Gustafsson, "Approximate bayesian smoothing with unknown process and measurement noise covariances," *IEEE Signal Processing Letters*, vol. 22, no. 12, pp. 2450–2454, Dec 2015.
- [3] Zoubin Ghahramani and Geoffrey E. Hinton, "Parameter estimation for linear dynamical systems," Tech. Rep., 1996.
- [4] E. zkan, F. Lindsten, C. Fritsche, and F. Gustafsson, "Recursive maximum likelihood identification of jump markov nonlinear systems," *IEEE Transactions on Signal Processing*, vol. 63, no. 3, pp. 754–765, Feb 2015.
- [5] C. Nemeth, P. Fearnhead, and L. Mihaylova, "Sequential monte carlo methods for state and parameter estimation in abruptly changing environments," *IEEE Transactions on Signal Processing*, vol. 62, no. 5, pp. 1245–1255, March 2014.
- [6] Yusuf Erol, Lei Li, Bharath Ramsundar, and Stuart J. Russell, "The extended parameter filter," in *Proceedings of the 30th International Conference on Machine learning*, 2013, The full version appeared as Tech. Rep. UCB/EECS-2013-48.
- [7] N. Chopin, P. E. Jacob, and O. Papaspiliopoulos, "Smc2: an efficient algorithm for sequential analysis of state space models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 75, no. 3, pp. 397–426, 2013.
- [8] V. Bastani, L. Marcenaro, and C. S. Regazzoni, "Online nonparametric bayesian activity mining and analysis from surveillance video," *IEEE Transactions on Image Processing*, vol. 25, no. 5, pp. 2089–2102, May 2016.
- [9] V. Bastani, L. Marcenaro, and C. Regazzoni, "A particle filter based sequential trajectory classifier for behavior analysis in video surveillance," in *Image Processing (ICIP), 2015 IEEE International Conference on*, Sept 2015, pp. 3690–3694.
- [10] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking," *IEEE Transactions on Signal Processing*, vol. 50, no. 2, pp. 174–188, Feb 2002.
- [11] James Hensman, Nicolo Fusi, and Neil Lawrence, "Gaussian processes for big data," in *Proceedings of the Twenty-Ninth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-13)*, Corvallis, Oregon, 2013, pp. 282–290, AUAI Press.
- [12] C.E. Rasmussen and C.K.I. Williams, *Gaussian Processes for Machine Learning*, MIT Press, Cambridge, MA, USA, 2006.
- [13] Michalis K. Titsias, "Variational learning of inducing variables in sparse gaussian processes," in *In Artificial Intelligence and Statistics 12*, 2009, pp. 567–574.
- [14] Andrew Mchutchon and Carl E. Rasmussen, "Gaussian process training with input noise," in *Advances in Neural Information Processing Systems 24*, J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, Eds., pp. 1341–1349, Curran Associates, Inc., 2011.
- [15] Roger Frigola, Yutian Chen, and Carl E. Rasmussen, "Variational Gaussian process state-space models," in *Advances in Neural Information Processing Systems 27 (NIPS)*, Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, Eds. 2014.
- [16] J. Q. Candela, A. Girard, J. Larsen, and C. E. Rasmussen, "Propagation of uncertainty in bayesian kernel models - application to multiple-step ahead forecasting," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, April 2003, vol. 2, pp. II–701–4 vol.2.